SN

The PSYCHOLOGICAL RECORD . . . .

NOVEMBER, 1940
Vol. IV No. 12

A STATISTICAL EVALUATION OF TESTS OF PERSISTENCE

DOROTHY RETHLINGSHAFER

THE PRINCIPIA PRESS, INC.
BLOOMINGTON, INDIANA

Price of this number, 25 cents

The Principia Press, Inc., has undertaken the publication of this co-operative journal to afford authors the opportunity of immediate publication at the least possible expense. The present low cost of publication and possible future reductions, depend entirely upon the number of subscriptions. The subscription price has been made low to induce individuals to subscribe. Under the Articles of Incorporation of the Principia Press no profit can be made on any of its publications. Therefore an increase in the number of subscribers will be reflected in reduced costs to authors and in increase in the number of pages published annually.

EDITORIAL BOARD. The above named board of associate editors have agreed to serve in an active capacity for a period of three years, and all manuscripts will be submitted to them.

MANUSCRIPTS may be sent to the appropriate associate editor or to Dr. J. R. Kantor. Longer papers (30 or more typewritten pages) in the above mentioned fields are especially desired.

COSTS. The cost of publication will be $2.00 per page of approximately 470 words. The charge for line cut illustrations will be $2.00 regardless of size. Special charges will have to be made for complex tabular matter and for half-tone or colored illustrations.

REPRINTS. One hundred copies of each paper will be supplied gratis. Additional copies may be purchased in any quantity at a rate which will be set when the author is informed of the cost of publication.

ROYALTIES. Fifty per cent of the net income from the sale of individual copies, or from copies sold as part of back volumes, will be credited as royalties to the author's account. Royalties cannot be paid on income from subscriptions current in the year of publication.

SUBSCRIPTIONS. The subscription price is $4.00 per year for an annual volume of approximately 500 pages. Individual papers will be sold at an advertised price, which will depend upon the size of the article. Foreign subscriptions will be $4.50.

CORRESPONDENCE concerning editorial matters should be addressed to Dr. J. R. Kantor, Indiana University. Business correspondence should be addressed to Dr. C. M. Louttit, Principia Press, Inc., Bloomington, Ind.

A STATISTICAL EVALUATION OF TESTS OF PERSISTENCE[1, 2]

By Dorothy Rethlingshafer
University of North Carolina

Introduction

One method of testing persistence is to determine the strength of a tendency-to-continue an activity once started by interrupting that activity and recording the behavior of the subjects following the interruption. This interruption technique has been used by various investigators, as (1), (2), (4), (5), (6), but no attempt has been made to differentiate subjects in the strength of their tendencies to continue. Results have been reported for the most part in terms of average percentage of resumption or non-resumption of the interrupted activities. Delayed resumptions, as contrasted to immediate returning to the blocked activities, have also been noted and percentage that they represented of the total records been recorded.

The author using the interruption technique in comparing normal and feebleminded subjects in the strength of their goal fixations has used a scale to measure the strength of the subject's tendency-to-continue. As previously described (7), it was found that the behavior following interruption could be classified into seventeen different categories ranging from strong evidence of tendency-to-continue—subjects completely refusing to be interrupted—to the behavior that would indicate none or little strength of tendency-to-continue—the subjects being easily interrupted and not returning to the original activity when given an opportunity to do so. These qualitative records of behavior arranged along a scale of tendency-to-continue were given quantitative scores by making the assumption that the distribution was normal and then determining the points on the scale in terms of

[1] Read in part before the Southern Society for Philosophy and Psychology, April 7, 1939, Durham, N. C.

[2] Recommended for publication by Dr. C. M. Louttit, October 2, 1940.

sigma units taken from a mean of zero (7). Complete-refusal-to-be-interrupted had a score of 2 in sigma units and at the other end of the scale, non-resumption was given the score of —1.2. Some of these seventeen categories could not be differentiated from each other when the scale points were kept to one decimal place, and the final scale that was used had twelve points, as is shown below.

Types of behavior following interruption.[3]	*Score value on scale*
A. Non-resumption	—1.2
B. Tendency to resume	— .8
C. Part-refusal-to-be-interrupted and non-resumption	— .7
D. 1. Part-refusal-to-be-interrupted and tendency-to-resume	— .6
2. Delayed resumption (3 to 5 minutes)	
E. Delayed resumption (1 to 3 minutes)	— .5
F. Delayed resumption (less than one minute)	— .2
G. Immediate resumption	.3
H. Part-refusal-to-be-interrupted and delayed resumption (2 to 5 minutes)	.7
I. Part-refusal-to-be-interrupted and delayed resumption (1 to 2 minutes)	.8
J. Part-refusal-to-be-interrupted and delayed resumption (less than one minute)	.9
K. Part-refusal-to-be-interrupted and immediate resumption	1.3
L. Complete-refusal-to-be-interrupted	2.0

By using these scale scores it was possible to examine statistically the results from the different tests with 58 subjects and to determine the best test or tests. It was hoped that the evaluation of these tests, which use the interruption technique, would lead to improved tests for further experimentation. Therefore to determine the accuracy of our measurement of tendency-to-continue, we applied the following statistical criteria to the eleven tests

[3] A more complete description of the behavior indicated on the scale from A to L above can be found in a previous article (7).

used: (1) the reliability of the battery, (2) a mean performance on an individual test that would indicate neither an extremely negative nor an extremely positive performance, for on the scale we used the mean was placed at zero, (3) the standard deviation of each test, (4) a measure of kurtosis (β_2—3 and σ of kurtosis), and a measure of skewness (β_1 and $\sigma\beta_1$), (5) the correlation between the test and the total battery, (6) the communality coefficient of the first largest factor obtained by the Thurstone Multiple Factor technique.

Description of the Tests

Eleven activities, as puzzles, modeling in clay, cutting patterns from paper, etc., were selected as the original activities to be interrupted on the basis of the following: (1) All of the tasks could be performed successfully by the subjects, and (2) all had definite goals stated in the directions. With some of the activities, substitutes were also presented following the interruption. Also in some of the activities, tendency-to-continue was further tested by raising barriers to the completion of the first activity as by destroying or by covering the interrupted material, or by removing the subject from the field. These interrupted activities, serving as tests of tendency-to-continue, had a standard method of presentation of the original and of the substitute activity as well as a standard method of interruption and of pressure placed upon any subjects who might continue.[4]

Subjects

The subjects were 29 feebleminded and 29 normal subjects of mental ages 6 to 9-11, and of chronological ages of 6-8 to 9-7 for the normal children and 11-7 to 22-11 for the feebleminded subjects. The intelligence quotients of the normal children were, of course, between 90 to 110 and all but one of the feebleminded were between 47 and 70. Inasmuch as the behavior of the two groups so closely resemble each other (7), (8), they were combined and this statistical analysis of the tests is made on the

[4] A detailed description of the exact procedure used with each test is on file at the University of North Carolina. The reader is also referred to (7).

basis of the 58 subjects. It should be remembered, however, that the tests are limited to a fairly narrow mental age range and are evaluated in terms of scores from two varying intelligent groups.

Application of the Statistical Criteria

(1). *Reliability of the Battery.* The subjects were first scored from the scale on their behavior following the interruption in the eleven different tests. By means of these score values each individual received a test score as well as combined score on the battery. If on one test the subject immediately yielded at the interruption but also immediately resumed, he received a score of .3. If on another test he refused in part to be interrupted but also immediately resumed, he received a higher score on tendency-to-continue as measured by that test, 1.3. Though it was impossible to obtain an estimate of the reliability of each test, coefficients for the entire battery were obtained by a formula derived by Kuder-Richardson (No. 8, p. 156, 3) giving .87 for the feebleminded group, .89 for the normals and .85 for the two groups. This formula was used instead of the more commonly employed Spearman-Brown formula because: (1) it gives an unique value; and (2) the assumptions are fewer than are necessary with the Spearman-Brown formula. Moreover "when the assumptions are rigidly fulfilled, the figures obtained are the exact values of test reliability . . . : if the assumptions are not met, the figures obtained are underestimates." (3, p. 159).

(2). *The mean performance of each test.* The mean on the normal scale was set at zero but as the scores on the scale given above can range from 2 to —1.2 and as each test was set up to determine the effect of some condition affecting tendency-to-continue, it is apparent that the means will necessarily vary on the scale according to the conditions of the tests. Those tests in which barriers were raised will have a lower mean performance, for instance, than when only substitutes are used after the interruption. Examination of the individual means shown in Table I indicates neither an extremely minus nor an extremely positive performance. The negative means most removed from zero were those of tests 8 and 9 where the barriers against resumption may

have been too strong. On the other hand test 11 with the most positive mean had such a high return to the original activity by all the subjects that it did not differentiate among them.

(3) and (4). Evaluating the tests in terms of the *standard deviations,* we would choose those tests which lead to the largest

TABLE I

CRITERIA APPLIED FOR THE EVALUATION OF TESTS OF PERSISTENCE

Tests		Mean	S. D.	β_1	β_2—3	*r_{it} Communality coefficient of largest factor after rotation	
No substitute used	Test 1	.34	.89	.273	— .825	.448	.028
Substitutes used	Test 2	.25	.75	—.308	— .625	.550	.239
	Test 3	.15	.88	.242	— .683	.640	.166
Similar Substitutes used	Test 4	.31	1.08	.022	—1.299	.605	.249
	Test 5	.29	.94	.072	—1.766	.716	.634
	Test 6	.26	.91	—.003	— .476	.769	.711
Substitutes used Also barriers present	Test 7	—.24	.90	.578	— .779	.707	.455
	Test 8	—.85	.62	1.70	1.89	.606	.366
	Test 9	—.49	.70	.807	— .316	.613	.181
Interruption by destruction of test material	Test 10	—.24	.54	—.651	— .822	.545	.319
Substitute used. Also other materials	Test 11	.45	.61	.711	1.015	.192	.011

* Correlation between test and total battery.

σ of β_1 .32
σ of β_2 .64

potential differentiation among subjects, eliminating tests on the basis of their small sigma values.

Likewise the poorer tests would tend to show *positive kurtosis*

whose peaked distribution would indicate that the test scores were too concentrated at one place on the scale, a condition which we found was likely to occur with these tests, due to the fairly small number of cases and the narrow range of possible scores. On both of the above two criteria, tests 8 and 11 would be eliminated.

In considering *skewness* we see that none of the tests have a value for β_1 large enough to have a critical ratio approaching three, except tests already criticized, 8, 9, and 11.

(5). *Correlation between test and battery.*[5] To the extent that our battery of eleven tests can be considered a standard against which the performance on the individual tests can be compared, we can evaluate the tests according to the size of the coefficients shown in the fifth column in Table I. Test 6 has the highest r of .77 and test 11 the lowest of any of the tests.

It is interesting to note certain characteristics about test 8 which has been criticized above but shows here a r of .60 with the total battery. Though compared to the other tests its mean was most removed from the zero point along the negative end of the scale of persistency, and though it had a low sigma value, a skewness of 1.70 and a kurtosis of 1.89, nevertheless this test correlates well with the total battery; moreover, we will find a first factor loading of .60. In this test a "strong" barrier was placed between the subjects and the original interrupted activity. This barrier effectively lowered resumption to such an extent that the scores all piled toward one end of the scale. Though the barrier was too strong to make this a "good" test of tendency-to-continue, in the sense that a normal distribution of scores would result, yet we see that this barrier was overcome only by those subjects who ranked high in "persistency" on the other tests, and was yielded to by those who possessed less tendency-to-continue. The whole distribution was moved down the scale and tended to pile up at that end, but the subjects maintained sufficiently their comparative rankings to continue to correlate fairly well in this test with the total battery.

(6). *Factors present in the tests.* The scale scores for the dif-

[5] All the correlations are spuriously high to the extent that each test is correlated with itself in the battery.

ferent subjects in the various tests can be assumed to result from the operation of various factors. By applying the Thurstone multiple-factor technique to the matrix of intercorrelations between the eleven tests, two centroid factors were obtained which when rotated into simple configuration gave one general factor with weights on all tests, except 1 and 11, of over .40, and a second factor with significant weights on only tests 1, 3, 9, and 11. The most general factor was named tendency-to-continue as representing that which the tests were set up to measure. The communality coefficient of this factor is given in the last column of Table I as a criterion of the validity of the tests. We see that test 6 ranks first when judged by this criterion while tests 1 and 11 are apparently not measuring tendency-to-continue.[6]

On the basis of the last two important criteria, test 6 is apparently the best measurement of what the tests were set up to measure, of what we have called tendency-to-continue. Moreover, it also ranks high according to the other indices we have used in evaluating the tests: its mean is not far removed from the zero; it ranks third in the size of its standard deviation; it is the least skewed of any of the tests; and its coefficient for kurtosis is negative.

Why test 6 represents the best test can be better understood by a comparison of it with test 11, the poorest test by all of the criteria. In test 6 the subjects were asked to cut from paper a pattern of an airplane—not an excellent pattern and one which was not frequently recognized as an airplane by the subjects. Because of the inferiority of the pattern, the material for this test might be called meaningless. Like nonsense material it carried less associations than the materials of the other tests. The experimenter, therefore, suggests that the less meaningful the more superior the test. Test 11, on the other hand, which was influenced the least by the factor named tendency-to-continue had the most "meaning" for the subjects; i. e., from the comments and

[6] We are not here concerned with the second factor as it was present in only four of the tests, none of which were found to be good tests. A tentative name of emotionality was given to the factor on the basis that the scores on the tests weighted with this factor may have been influenced by timidity, uncertainty, or distrust of ability on the part of the subjects.

behavior of the subjects when in the presence of the large blocks, which was the material for test 11, the experimenter concluded that these blocks "appealed" more than any of the other materials used. Since there was such a strong "appeal," the test as a measure of tendency-to-continue no longer functioned. The subjects resumed frequently enough; they continued this activity in spite of the interruption; but their comparative positions in scale scores as measured by the other tests, were not maintained; hence this test correlated less with the total battery than any other of the measures used.

In contrast to this in test 8, where as has been pointed out there was also a piling up of the scores, though at the negative end of the scale because of a strong barrier against resumption, the subjects maintained sufficiently their rankings that this test correlated well with the battery and had a first factor loading of .61. Evidently the piling up of the scores at one end of the scale merely indicated that the barrier was working, but did not imply that tendency-to-continue was not being measured.

Application of Further Criteria to the Tests

Additional standards for the evaluation of tests are the ease of administration, i. e., in these tests the ease with which a standard procedure could be followed in the presentation of the original activity, and in the accuracy with which a fixed place of interruption could be attained. It was found most easy to impose a standard method of presentation and of interruption with blocks, puzzles and least easy with modeling clay figures which permitted slight differences in the way that the directions were given from subject to subject. However, an examination of the tests results to determine if any differences between them could be accounted for by slight differences in degree of standardization brought negative results; some of the poorer tests, as test 11, were excellently standardized. Tests 3, 4, and 9, representing tasks which demanded the modeling of clay figures, do not rank high according to the statistical criteria, but they also are not the lowest in their rankings. The experimenter is of the opinion that the standardization of procedure was carried sufficiently far with all the tests so that any differences among them cannot be accounted for by variance in standardization.

Conclusion

One of the many methods that has been employed to measure persistence is to block the subject from continuing some activity he has started. Eleven tests using this interruption technique were used with 58 subjects of mental ages 6 to 9-11. Scores were given these subjects in terms of their behavior following the interruption. An evaluation was then made of these eleven tests for tendency-to-continue by considering the following criteria: (1) the reliability of the battery; (2) a mean close to the zero point on the normal scale; (3) the size of the standard deviation; (4) coefficients of kurtosis and skewness; (5) the correlation between each test and the total battery; (6) the communality coefficient of the first largest factor after rotation; and (7) standardization of procedure.

According to the above criteria the most superior test was one in which the material could be described as meaningless. Like nonsense material it carried fewer associations for the subjects than the materials used with some of the other tasks. It was freer from the factor of interest in the material *per se*, and was, therefore, a purer test of a general tendency-to-continue an activity once started.

BIBLIOGRAPHY

1. Adler, D. L., & Kounin, J. S. Some factors operating at the moment of resumption of interrupted tasks. *The Journal of Psychology*, 1939, *7*, 255-267.
2. Köpke, Paul. Substitute satisfaction with normal and feeble-minded children (Unpublished study). Reported by Lewin, K. *Dynamic Theory of Personality*, New York, McGraw-Hill Book Co., Inc., 1935, 202-204.
3. Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability, *Psychometrika*, Vol. 2, No. 3, September 1937, 151-160.
4. Lissner, K. Die Entspannung von Bedürfnissen durch Ersatzhandlungen. *Psychol. Forsch.*, 1933, *18*, 218-250.
5. Mahler, W. Ersatzhandlungen verschiedenen Realitätsgrades. *Psychol. Forsch.*, 1933, *18*, 27-89.
6. Ovsiankina, M. Wiederaufnahm unterbrochener Handlungen, *Psychol. Forsch.*, 1928, *11*, 302-389.

7. Rethlingshafer, Dorothy. Measures of Tendency-to-continue: I. Behavior of feebleminded and normal subjects following the interruption of activities. *Journ. of Genet. Psychol., In press.*
8. Rethlingshafer, Dorothy. Measures of Tendency-to-continue: II. Comparison of feebleminded and normal subjects when interrupted under different conditions. *Journ. of Genet. Psychol., In press.*